PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C07K

A2

(11) International Publication Number: **WO 99/62930**

(43) International Publication Date: 9 December 1999 (09.12.99)

(21) International Application Number: PCT/US99/12221

(22) International Filing Date: 2 June 1999 (02.06.99)

(30) Priority Data:

60/087,785	3 June 1998 (03.06.98)	US
Not furnished	1 June 1999 (01.06.99)	US

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(71) Applicant: MILLENNIUM PHARMACEUTICALS, INC. [US/US]; 75 Sidney Street, Cambridge, MA 02139 (US).

(72) Inventor: DANCIK, Vladimir; 4 Foyal Crest Drive #3, North Andover, MA 01845 (US).

(74) Agents: HANLEY, Elizabeth, A. et al.; Lahive & Cockfield, LLP, 28 State Street, Boston, MA 02109 (US).

(54) Title: PROTEIN SEQUENCING USING TANDEM MASS SPECTROSCOPY

(57) Abstract

A new algorithm, SHERENGA, for de novo spectral interpretation is described that automatically learns fragment ion–types and intensity thresholds from a collection of test spectra generated from any type of mass spectrometer. The algorithm employs a graph theory approach. The test data is used to construct optimal path scoring in the graph representations of tandem mass spectra. A ranked list of high scoring paths corresponds to potential peptide sequences. SHERENGA is most useful for interpreting sequences of peptides resulting from unknown proteins not yet encountered in genome sequencing, and leveraging text based pattern matching for homology matching to known proteins. The algorithm also serves as a powerful adjunct for validating the results of database matching algorithms in fully automated, high–throughput peptide sequencing.

PROTEIN SEQUENCING USING TANDEM MASS SPECTROSCOPY

Background of the Invention

In a few seconds, a tandem mass spectrometer is capable of automatically ionizing a mixture of peptides and measuring their respective parent mass/charge ratios, then selectively fragmenting each peptide into constitutive pieces and measuring the mass/charge ratios of the fragment ions (MS/MS spectra of peptides). The peptide sequencing problem is then to derive the sequences of peptides given their MS/MS spectra. For an "ideal" fragmentation process and an "ideal" mass-spectrometer the sequence of the peptide could be simply determined by converting the mass differences of consecutive fragmentions in the spectrum to their corresponding amino acids. In practice, de novo peptide sequencing remains an open problem and even simple spectrum may require tens of minutes for a trained expert to interpret.

The previous attempts to develop automated de novo peptide sequencing algorithms followed either global or local search paradigms. One prior approach involves the generation of all amino acid sequences and corresponding electronic spectra , i.e. calculation of all theoretically possible fragment masses for each sequence. The goal is to find a sequence with the best match between the experimental and electronic spectrum. Since the number of sequence permutations grows exponentially with the length of the peptide, different pruning techniques were designed to limit the combinatorial explosion in global methods. Prefix pruning restricts the computational space to sequences whose prefixes match the experimental spectrum well. Unfortunately, prefix pruning frequently discards the correct sequence if its prefixes are poorly represented in the spectrum. The number of sequence permutations examined can be further pruned by limiting the possible amino acid composition derived either through chemical amino acid analysis or through composition measurement for ions below m/z 160 in the tandem mass spectrum. The difficulty with the prefix approach is that pruning frequently discards the correct sequence if its prefixes are poorly represented in the spectrum. Another intrinsic problem with the global approach is that the spectrum information is used for scoring only after the potential peptide sequences are generated.

The global approach de novo programs typically have running time on the order of hours.

Local approaches tend to be more efficient techniques for de novo peptide sequencing because they use the spectral information before any candidate sequence is evaluated. In different modifications of the local approach the fragment ions correspond (sometimes implicitly) to vertices of the spectrum graph as described in, "Christian Bartels. Fast algorithm for peptide sequencing by mass spectroscopy" 19:363-368, 1990 which is incorporated by reference herein.

The peaks in the spectrum serve as vertices in the spectrum graph while the edges of the graph correspond to linking of vertices differing by the mass of an amino acid residue. Fundamental to graph theory approaches is the prior transformation of each peak in the experimental spectrum into several vertices in a spectrum graph. Each vertex represents a different possible fragment ion type assignment for the peak. The de novo peptide sequencing problem is thus cast as finding the longest path in the resulting directed acyclic graph. Since the number of edges in the spectrum graph is at most quadratic in the number of ions in the spectrum and since efficient algorithms for finding the longest paths are known such approaches have the potential to efficiently prune the set of all peptides to the set of high-scoring paths in the spectrum graph.

Although de novo sequencing software programs were developed beginning in the late 1980's, none are in widespread use today. The more widely used database search programs rely on the ability "to look the answer up in the back of the book" when studying genomes of extensively sequenced organisms. While de novo interpretation is limited to a certain extent by ambiguities arising from incomplete fragmentation of a peptide in a tandem mass spectrometer, current de novo algorithms implementations of graph-theory approaches face the following unsolved computational problems.

Existing algorithms tend to be instrument-dependent, i.e. they are designed for the kind of fragment ions that are most likely for the authors' particular type of mass spectrometer. No rigorous approach to defining ion-types and intensity thresholds in an instrument-independent fashion has yet been proposed.

If the peptide fragmentation is incomplete the spectrum graph may break into a number of disconnected components. Random noise in the spectrum may generate many false vertices and edges in the spectrum graph that can mimic the correct peptide in the absence of a good scoring schema. Errors in the parent mass/charge assignment lead to misalignment between N-terminal and C-terminal vertices in the spectrum graph. No computational approach to adjust inappropriate parent mass/charge assignment has yet been proposed.

No rigorous approach to scoring paths in the spectrum graph has yet been proposed.

The longest path in the spectrum graph may correspond to unrealistic solutions because it uses multiple graph vertices associated with the same spectral peak (anti-symmetric path problem). No approach to take into account internal fragment ion types has yet been proposed. No approach to analyze ions of unknown charge state has yet been proposed.

High-throughput peptide sequencing via tandem mass spectrometry (MS/MS) is emerging as one of the most powerful tools in proteomics for identifying proteins. While de novo MS/MS peptide sequencing remains a difficult problem, our method, as implemented in SHERENGA Software, is not limited to the near-complete sequences contained in spectra generated on magnetic sector instruments employing high-energy collision induced dissociation. A new method implemented in software for de novo peptide sequencing by tandem mass-spectrometry is desirable. Our algorithm automatically learns ion-types, error rates and intensity thresholds from a collection of spectra.

Detailed Description of an Illustrative Embodiment

Let A be the set of amino acids with molecular masses $w(a)$, $a \in A$. A (parent) peptide $P = p_1,...p_n$ is a sequence of amino acids, the mass of peptide P is $m(P)=\Sigma\, m(p_i)$. A partial peptide $P' \subset P$ is a substring $p_i,...p_j$ of P of mass $\Sigma_{i \le t \le j}\, m(p_t)$. Electronic spectrum E(P) of peptide P is a set of masses of its partial peptides. An (experimental) spectrum $S=\{s_1,...s_m\}$ is a set of masses of (fragment) ions. . A mass s

matches a peptide P if m(P')=s for a partial peptide P'⊂P .. Denote x(s,P)= 1 if s matches P, and x(s,P)=0 otherwise. A match $m(S,P) = \Sigma_{s \in S} m(s,P)$ between spectrum S and peptide P is the number of ions from the spectrum S that match peptide P. In another words, m(S,P) is the number of masses that experimental and electronic spectra have in common.

The peptide sequencing problem can stated as follows. Given spectrum S and a parent mass m find a peptide of mass m with the maximal match to spectrum S.

However, different mass-spectrometers lead to different variations of the peptide sequencing problem. In particular, peptide fragmentation in a tandem mass-spectrometer is characterized by a set of numbers $\Delta=\{\delta_1,..., \delta_k\}$ called ion-types. A δ-ion of a partial peptide $P' \subset P$ is such modification of P' that has molecular mass m(P')-δ. In this case, a molecular mass s matches a peptide P if w(P')-δ=s for a partial peptide $P' \subset P$ and an ion-type $\delta \in \Delta$. For tandem mass-spectrometry, electronic spectrum E of peptide P is created by subtracting all offsets from Δ from the masses of all partial peptides of P (denoted as $E \Theta \Delta$).

The problem can be further stated as given spectrum S, the set of ion types Δ, and the mass w, find a peptide of mass w with the maximal match to spectrum S.

Denote partial N-terminal peptide sequences $p_1,...,p_i$ as P_i , i = 1,....,n-1 and partial C-terminal peptide sequences $p_j,....,p_n$ as P_j^- , j = 2,...,n (Note that this indexing differs from usual indexing for C-terminal ions.). In tandem mass-spectrometry spectrum S consists mainly of δ-ions of partial N-terminal and C-terminal peptides with δ being limited to a set of ion types $\Delta = \{\delta_1,....,\delta_k\}$. For example, the most frequent N-terminal ions are b, a, b-H_2O, b-NH_3 ~ ($\Delta=\{1,-27,17,16\}$) for out mass-spectrometer.

For a given partial peptide P_i let $W_\Delta(P_i)$ be the set of mass of all δ-ions of P_i for $\delta \in \Delta$, i.e. $W_\Delta(P_i) = \{m(P_i)-\delta_1....,m(P_i)-\delta_k\}$ For a peptide P we set $W_\Delta(P)= W_\Delta(P_1) \cup ...\cup W_\Delta(P_{(n-1)})$.

The mass spectrometry reconstruction problem then can be formulated as follows. For a given molecular mass w, spectrum S and the set of ion types Δ find a peptide P such that m(P) = w and $W_\Delta(P) = S$. We realistically try to find a peptide with the best match between spectrum and m(P), i.e. to maximize the size of W(P)\cap S

Assume, that a spectrum from a tandem mass-spectrometer consists mainly of N-terminal ions. and random noise. The correspondence between elements of spectra and vertices of spectrum graphs is closely tied with possible ion types. We capture the relationship between spectrum S and ion types $\Delta=\{\delta_1,....,\delta_k\}$ by a spectrum graph $G_\Delta(S)$. Vertices of the graph are integers representing potential masses of partial peptides, for vertex v we denote this mass by m(v). Every peak of spectrum $s\in S$ generates k vertices $V(s) = \{s+\delta_1,...., s+\delta_k\}$, with $m(v_i)=s+\delta_i$, i=1,....,k.

The set of vertices of spectrum graph then is $\{s_{\{initial\}}\} \cup V(s_1)\cup....\cup V(s_m)\cup \{s_{\{final\}}\}$, where $\{s_{\{initial\}}\} = 0$ and $\{s_{\{final\}}\}= m(P)$. Two vertices u and v are connected by a directed edge from u to v if v-u is the mass of some amino acid and the edge is labeled by this amino acid. If we look on vertices as potential partial N-terminal peptides, the edge from u to v implies that the sequence at v may be obtained by extending the sequence at u by one amino acid.. A spectrum S of a peptide P is called "complete" if S contains an ion corresponding to P_i for every $1 \le i \le n$. The use of spectrum graph is based on the observation that for a complete spectrum S of peptide P, S is a complete spectrum of a peptide P when there exists a path of length n from $v_{\{initial\}}$ to $v_{\{final\}}$ in $G_\Delta(S)$ that is labeled by P and $|W(P) S| = \Sigma_{v\in t}\ s(v)$, there s(v) denotes the multiplicity with which vertex v was created.

This observation transforms the tandem mass-spectrometry protein sequencing problem into finding the correct path in the set of all paths. Since the number of paths in the graph is enormous, we need some way of evaluating the paths. Previous implementations of the spectrum graph searched for a path visiting as many vertices as possible. Unfortunately, experimental spectra are frequently incomplete and noisy, i.e. they contain many peaks that do not correspond to any ions. Thus in order to find 9 peptide sequence corresponding to the given spectrum we have to develop a new approach to spectrum graph and scoring schema to deal with incomplete noisy spectra

and to evaluate the weight of the paths in the spectrum graph. Another problem is that different mass-spectrometers have different characteristics and different ion-types and therefore every algorithm for de novo peptide sequencing should be adjusted for a particular type of a mass-spectrometer. To address this problem an offset frequency function is described and an algorithm for an automatic, learning of ion types and intensity thresholds, and scoring parameters from a sample of experimental spectra is described.

An offset frequency function is introduced that represents an important new tool for defining the ion type tendencies for particular mass-spectrometers. The offset frequency function allows one to compare different mass spectrometers based on their propensity to generate different ion types thus making our algorithm instrument-independent.

Another observation is that in spectra we often observe ions that correspond to partial sequences $P_j^- = p_j \ldots p_n$ (these are called C-terminal ion as opposed to N-terminal ions corresponding to p_i's) and in some cases internal partial sequences $P_{\{ij\}} = p_i \ldots p_j$.

We distinguish between the name of ion type δ and the mass difference (offset) δ corresponding to ion type δ.

Consider a spectrum corresponding to peptide P from the learning sample. We will concentrate on peaks of spectra that are close to the mass of a partial peptide P_i. Peaks in a spectrum either represent random noise or δ-ions of partial peptides.

If we don't know the ion types $\Delta=\{\delta_1,\ldots,\delta_k\}$ produced by a given mass spectrometer we cannot interpret the spectrum. We must distinguish between random noise and δ-ions. We describe how to learn the set Δ and ion propensities from a sample of experimental spectra.

Let $S=\{s_1,\ldots,s_m\}$ be a spectrum corresponding to the peptide P and let $d(S) = \frac{m(P)}{m}$ be the average distance between the peaks. A partial peptide P_i and a peak s_j have an offset $x_{\{ij\}} = m(Pi)-sj$; for illustration purposes we shall treat x_{ij} as a discrete random variable. Given an arbitrary offset x, the probability that there is $s_j \in S$ such that

$x_{ij} = x$ can be roughly estimated as $\frac{1}{d(S)}$. For an offset $\delta \in \Delta$ the probability that there is a peak in S with offset $x_{ij} = \delta$ is approximately $(1-p(\delta))\frac{1}{d(S)} + p(\delta)$ where $p(\delta)$ is the probability of δ-ion (the portion of partial peptides that produce δ-ions). For example, the average d(S) for our sample spectra is 17.5, therefore probability of random offset is 0.057. The probability of an a-ion with offset -27 is 0.23. Thus the offset -27 is observed 4 times more frequently then the average offset. The statistics of offsets over all ions and all partial peptides provides a reliable learning algorithm for ion types.

Given spectrum S, offset x and precision ε we compute the number H(x,S) of pairs (P_i,s_j), $i = 1,...,n-1$, $j = 1,...,m$ that have offset $m(P_i)-s_j$ within distance from x. The offset frequency function is defined as $H(x) = \sum_S H(x,S)$, where the sum is taken over all spectra from the learning sample. To learn about C-terminal ions we do the same for pairs (P_i^-,s_j). Fig. 1 presents the plots of function H(x) for N-terminal, C-terminal, internal and doubly charged ion types. We consider only offsets within interval (-m,m) where m is the mass of the lightest amino acid. Vertical axes represent normalized offset counts with 1 being the average count Offset increment = 0.2. The only significant offsets for internal ions correspond to b and b--H_2O ions. The only significant offsets for doubly charged ions correspond to y and y--H_2O ions.

Offsets $\Delta = \{\delta_1,..., \delta_k\}$ corresponding to peaks of H(x) represent the ion-types produced by a given mass-spectrometer. Under normal circumstances we expect these offsets to correspond to the ion types that have sufficient support by chemistry.

TABLE 1

Offset	Integer Offset	Count	Filtered Count	Probability	Average Intensity	Term	Ion
18.85	19	604	604	0.6895	4.5457	C	y
0.85	1	568	565	0.6484	2.2788	N	b
-17.05	-17	338	338	0.3858	1.1966	N	b-H_2O
0.90	1	248	231	0.2831	0.5716	C	y-H_2O
-27.15	-27	204	200	0.2329	0.7537	N	a
20.05	20	183	180	0.2089	3.4699	C	y^2
-16.15	-16	159	152	0.1815	1.2766	N	b-NH_3
1.90	2	131	114	0.1495	0.6680	C	y-NH_3
-35.20	-35	151	141	0.1724	0.5253	N	b-H_2O-H_2O
-34.20	-34	134	131	0.1530	0.4736	N	b-H_2O-NH_3
-44.25	-44	129	126	0.1473	0.5516	N	a-NH_3
-45.15	-45	107	98	0.1221	0.4820	N	a-H_2O
2.30	2	102	95	0.1164	1.7460	C	y^2-H_2O
-16.10	-16	97	84	0.1107	0.4913	C	y-H_2O-NH_3
-17.15	-17	91	71	0.1039	0.4935	C	y-H_2O-H_2O

Table 1: Information about terminal ion types learned from experimental spectra. The remaining offsets have average count 45 and average intensity 0.431024. When computing filtered counts, the peaks that have been identified as ions are not counted again for subsequent ion types.

Table 1 contains the list of offsets that have larger than expected counts and the corresponding ion types as known in chemistry All the significant offsets we found correspond to known ion types Surprisingly enough, some ion types turned to be more significant than previously thought (i.e. b-H_2O-H_2O has larger count that y-NH_3). Also Fig. 1 clearly shows the presence of internal b-ions in the spectra.

A part of the learning of ion types is to decide what interval of offsets should be considered for particular ion type. The error range is chosen to be the width of the corresponding peak in the plot of the offset frequency function of H(x). This analysis suggests that $\varepsilon = 0.45$. If we need to be more precise, we can assume that offsets are

distributed according to the mixture of uniform and normal distributions and use maximum likelihood methods to estimate appropriate values for error ranges.

Once we have learned and selected significant ion types we can annotate spectra from our learning sample. Annotated spectra will provide support for learning other features needed for the construction of spectrum graphs.

Peaks in a spectrum differ in intensity and one has to address the question of setting a threshold for distinguishing the signal from noise in a spectrum prior to transforming it to a spectrum graph. Low thresholds lead to excessive growth of the spectrum graph while high thresholds lead to fragmentation of the spectrum graph. Earlier de novo sequencing algorithms set up the intensity thresholds for experimental spectra in a largely heuristic manner and have not addressed the fact that the intensity thresholds are ion-type dependent. The offset frequency function allows one to set up intensity thresholds in a rigorous way.

Given a spectrum, we can address this concern by normalizing and ranking group intensities into bins of size K and rank K peaks with largest intensity by 1, next K peaks are ranked by 2 and so on. A natural choice for K is the length of the underlying peptide. Since this information is usually unavailable, K may be chosen as the ratio of the peptide mass and the average mass of an amino acid. We normalize intensities of peaks in a spectrum in such way that the average intensity of the peaks in the spectrum is 1. The frequences of ion types depending on intensity are shown in Figure 3.

The change of H(x) depending on the intensity rank is shown in Fig. 2, which guides us in selecting intensity thresholds. Fig. 2 convincingly demonstrates that the intensities ranked below 5 represent nothing but random noise since the offset frequency function has no pronounced peaks in this region. It implies that for an average MS/MS spectrum on an ion-trap instrument no more than about 60 top intensities should be considered as a potential signal. This observation suggests a limit for the number of peaks analyzed by any peptide MS/MS interpretation program and indicates that the analysis of 100+ peaks with any program is likely to hamper rather than to help in interpreting peptide sequences.

Moreover, Fig. 3 demonstrates that intensity thresholds are ion-type dependent. For example, the analysis of b-ions can be limited to intensity ranks 1, 2 and 3, while the analysis of b-H_2O can be limited to intensity ranks 3, 4 and 5. A similar analysis implies that only intensities ranked 1 and 2 (i.e 20-30 high-intensity peaks) should be considered for y-ions while intensities ranked 2, 3 and 4 represent potential y-H_2O ions.

For example, Fig. 3 shows that only intensities ranked 1 and 2 should be considered for y-ions while intensities ranked 2, 3 and 4 represent potential y-H_2O ions.

The approach to construction of the spectrum graph described above is incomplete since it does not take into account inaccuracies in experimental mass measurements of fragment and parent ions. Let partial peptide P_i produces peaks $s_1,...,s_k$ in the spectrum corresponding to the ion types $\delta_1,...,\delta_k$. Above we assumed that $s_1 + \delta_1 = s_2+\delta_2 = ... = s_k + \delta_k = m(P_i)$ and all k ion types generate the same vertex in the spectrum graph. Of course, this is not the case for real spectra. Due to inaccuracies of experimental mass measurements the peaks $s_1,...,s_k$ correspond to different vertices with weights $s_j + \delta_j$, $1 \le j \le k$ within mass tolerance that is instrument dependent.

The merging algorithm decides what vertices in the spectrum graph are to be merged into one vertex. It is important to merge appropriate vertices; if we do not merge vertices that correspond to the same partial peptide, we will interpret meaningful peaks of spectra as a noise. On the other hand, if we merge vertices that do not correspond to the same peptide, we may interpret noise as meaningful peaks. To address this problem SHERENGA uses greedy a algorithm for merging vertices and introduces bridge edges in the resulting graph.

If a peptide undergoes incomplete fragmentation the spectrum graph does not contain a vertex corresponding to an underrepresented position in a peptide. Since fragmentation is frequently incomplete many peptides contain positions that have no corresponding peaks in the spectra. This can lead to a fragmented graph or, more frequently, a graph with paths that do not correspond to feasible solutions. This effect only amplifies as we introduce thresholds and exclude low intensity peaks from the spectrum. To overcome this problem we modify the spectrum graph by introducing gap edges. A gap edge in the spectrum graph is a directed edge from u to v such that v - u is

the mass of a dipeptide, i.e. the sum of masses of two amino acids. In a more general approach we consider tri-peptides or even longer peptides.

Accurate determination of the peptide parent mass/charge is extremely important in de novo peptide sequencing. An error in parent mass leads to systematic errors in the masses of vertices for C-terminal ions thus making peptide reconstruction difficult. In practice, the offsets between the real peptide masses (given by the sum of amino acids of a peptide) and experimentally observed parent mass/charge as shown in Fig. 4 are frequently so large that the errors in peptide reconstruction become almost unavoidable. To address this problem we have designed a combinatorial algorithm for parent mass/charge computation that provides a more accurate determination of the parent mass.

The goal of scoring is to answer the question of how well a candidate peptide "explains" a spectrum and to choose the peptide that explains the spectrum the best. Below we introduce a probabilistic model for tandem mass-spectrometry and derive a rigorous scoring algorithm (versus largely heuristic previous approaches).

Let p(P,S) be the probability that a spectrum S is generated by a peptide P produces spectrum S. It is appropriate to design scoring schema so that the high scoring peptides P have the high probability p(P,S). Below we describe a probabilistic model, evaluate p(P,S) and derive a scoring schema for paths in the spectrum graph. by the probabilities of the responding peptides. The longest path in the weighted spectrum graph corresponds to the peptide P that "explains" spectrum S the best.

In a probabilistic approach tandem mass spectrometry is characterized by a set of ion types $\Delta = \{\delta_1,...,\delta_k\}$ and their probabilities $\{p(\delta_1),..., p(\delta_k)\}$ such that δ_i-ions of a partial peptide $P' \subset P$ are produced independently with probabilities $p(\delta_i)$. A mass-spectrometer also produces a "random noise'" that in any position may generate a peak with probability q_R. Therefore, a peak at position corresponding to a δ_j-ion is generated with probability $q_j = p(\delta_j) + (1-p(\delta_j))q_R$ that can be estimated from the observed empirical distributions (Table 1). Partial peptide P' may theoretically have up to k corresponding peaks in the spectra. It has all k peaks with probability $\prod_{i=1}^{k} qi$ and it has no peaks with probability $\prod_{i=1}^{k}(1-qi)$.

The described probabilistic model defines probability p(P,S) that a peptide P produces spectrum S . We can formulate peptide sequencing problem now as follows:

For a given spectrum S find a peptide P maximizing p(P,S), i.e. $p(P,S) = \max_P p(P,S)$.

To illustrate the idea of scoring informally let's assume that only 4 types of ions are possible: y, b, y-H_2O, b-H_2O ions with probabilities of appearing q_1,q_2,q_3,q_4 . Assume also that probability of random noise is q_R.

Suppose that a candidate partial peptide P_i produces ions δ_1, ..., δ_l ("present" ions) and does not produce the ions δ_{l+1},..., δ_k ("missing" ions) in the spectrum S . These l "present" ions will result in a vertex in the spectrum graph corresponding to P_i. How should we score this vertex? The existing database search algorithms use "premium for present ions" approach suggesting that the score for this vertex should be proportional to q_1 --- q_l or maybe $\frac{q_1}{q_R} \cdots \frac{q_l}{q_R}$ to normalize the probabilities against the noise. (The ratios $\frac{q_l}{q_R}$ can be taken from the offset frequency function). Normalizing against the noise has the additional effect of penalizing peaks of the experimental spectrum that are not explained in relation to a candidate sequence. Below we also show that it is not a correct approach and that we have to. However we achieve better results when we significant improvement results from penalizing for non-presence of ions in the experimental spectrum which are possible from fragmentation of a candidate sequence. The probability score of the vertex is then given by

$$\frac{q_1}{q_R} \cdots \frac{q_l}{q_R} \frac{(1-q_{l+1})}{(1-q_R)} \cdots \frac{(1-q_k)}{(1-q_R)}$$

("premium for present ions, penalty for missing ions"). This important observation was overlooked in scoring the database search hits for peptide mass-spectrometry. Although "premium for present ions, penalty for missing ions" approach may sound counter-intuitive, it is confirmed both by our theoretical analysis and improvements in SHERENGA performance as compared to the previous approach.

We explain the role of this principle for a resolution of a simple alternative between dipeptide GG and amino acid N of the same mass. In the absence of "penalty for missing ions" GG is selected against N in the presence of any (even very weak random noise) peak supporting the position of the first G. Our results implies that such a rule leads to many wrong GG-abundant predictions since our learning procedure implies that the weak peak after the first G is, in fact, a vote against GG. The correct rule is to vote for GG if it is supported by a b or y-type ion. This rule is automatically enforced by our "premium for present ions, penalty for missing ions" scoring. The same concepts extend to ambiguities between AG and GA vs. K or Q (all mass 128).

For the sake of simplicity we assume that all partial peptides P_i and P_i^- are equally likely and ignore the intensities of peaks for now. We discretize the space of all masses in the interval from 0 to the parent mass m(P)=M, denote T = {0,..., M}, and represent the spectrum as an M-mer vector $S = \{s_1,...,s_M\}$ such that s_t, is the indicator of resence/absence of peaks at position t ($s_t = 1$ if there is a peak at position t and $s_t = 0$ otherwise). For a given peptide P and position t, s_t is a 0-1 random variable with probability distribution $p(P,s_t)$. For a given P probabilities $p(P,s_t)$ are independent and

$$p(P,S) = \prod_{t=1}^{M} p(P,s_t).$$

Let $T_i = \{t_{i1},...,t_{ik}\}$ be the set of positions that represent Δ-ions of a partial peptide P_i where $\Delta = \{\delta_1,...,\delta_k\}$. Let $R=T \cup T$ be the set of positions that are not associated with any partial peptides. The probability distribution $p(P,s_t)$ depends on whether $t \in T_i$ or $t \in R$. For a position $t = t_{ij} \in T_i$ the probability $p(P,s_t)$ is given by

$$p(P,s_t) = \{q_j, \text{ if } s_t - 1 \text{ (i.e. a peak is generated at position t) and } 1 - q_j, \text{ otherwise}$$

Similarly for $t \in R$ the probability $p(P,s_t)$ is given by

$$P_R(P,s_t) = \{ q_R, \text{ if } s_t = 1 \text{ (i.e. there is a random noise at position t), and } 1-q_R, \text{ otherwise}$$

and the overall probability of 'noisy' peaks in the spectrum can be estimated as

$$\prod_{t \in R} p_R(P,s_t).$$

Let $p(P_i,S) = \prod_{t \in Ti} p(P,s_t)$ be the probability that a peptide P_i P_i^+ and P_i^- produces a given spectrum at positions from the set T_i (all other positions ignored). For the sake of simplicity, assume that each peak of the spectrum belongs only to one set T_i and that all positions are independent. Then

$$p(P,S) = \prod_{t=1}^{M} p(P,s_t) = \left(\prod_{i=1}^{n} p(P_i,S) \right) \prod_{t \in R} P_R(P,s_t).$$

We also assume that all positions from R have the same probability distribution $p_R(t)$ independent of P. For a given spectrum S the value $\prod_{t \in T} p_R(P,s_t)$ does not depend on P and the maximization of p(P,S) is the same as the maximization of

$$\frac{p(P,S)}{p_R(S)} = \frac{\prod_{i=1}^{n} \prod_{j=1}^{k} p(P,s_{tij}) \prod_{t \in R} p_R(P,s_t)}{\prod_{t \in R} p_R(P,s_t)} = \prod_{i=1}^{n} \prod_{j=1}^{k} \frac{p(P,s_{tij})}{p_R(P,s_{tij})}$$

In logarithmic scale the above formula together with 1 and 2 imply the additive "premium for present ions, penalty for missing ions" scoring of vertices in the spectrum graph.

Although we explain our approach in the terms of probability, the calculations are done in logarithmic scale to avoid dealing with very small numbers that may lead to the loss of precision. Up to this point we ignored the intensities of the peaks and the scoring described above assigns the same score to low intensity and high intensity peaks. To incorporate the intensities into scoring we assume that intensity for ion type δ_j is distributed according to empirical distribution $I_{\delta i}(x)$ and modify formulas (1) and (2) accordingly.

The protein sequencing algorithm involves the generation of the weighted spectrum graph (as described above) and the search for the highest scoring paths in the spectrum graph.

After the weighted spectrum graph is constructed we cast peptide sequencing problem as the longest path problem in directed acyclic graph. This problem is solved by a fast linear time dynamic programming algorithm with running time O(E), where E is the number of edges in the spectrum graph. For a typical spectrum, the algorithm is

very fast thus giving the spectrum graph approach an advantage over the global approaches.

Unfortunately, this simple algorithm does not quite work in practice. The problem is that every peak in the spectrum may be interpreted either as an N-terminal ion or C-terminal ion. Therefore, every "real" vertex (corresponding to a mass m) has a "fake" twin vertex (corresponding to a mass m(P)-m-offset). Moreover, if the real vertex has a high score then its fake twin also has a high score. The longest path in the spectrum graph then tends to include both real vertex and its fake twin since they both have high scores. Such paths do not correspond to feasible protein reconstructions and should be avoided. However, the known longest path algorithms do not allow to avoid such paths. Since they cannot check back on whether one of the twins was already included in the growing path. This problem was overlooked in the previous work on de novo protein reconstruction.

Therefore, the simple reduction of the tandem mass-spectrometry peptide sequencing to the longest path problem described earlier is inadequate. We now describe the anti-symmetric longest path problem that adequately models the peptide sequence reconstruction.

Let G be a graph and let T be a set of forbidden pairs of vertices of G (twins). A path in G is called anti-symmetric if it contains at most one vertex from every forbidden pair. Anti-symmetric longest path problem is to find a longest anti-symmetric path in G with a set of forbidden pairs T.

The intrinsic property of the conventional longest path algorithms is that they use only neighbors of a given vertex while computing the shortest path ending in this vertex. Since vertices in a forbidden pair are not necessarily neighbors, these algorithms can not be adjusted to find anti-symmetric longest paths. The anti-symmetric longest path problem is NP-hard thus indicating that efficient algorithms for solving this problem are unlikely.

This negative result does not imply yet that it is futile to attempt to find an efficient algorithm for tandem mass-spectrometry peptide sequencing since this problem has a special structure consisting of forbidden pairs that leads to an efficient algorithm

for finding anti-symmetric longest paths. Below we show that it is exactly the case and design an efficient algorithm for the tandem mass-spectrometry problem.

Vertices in the spectrum graph are numbers that correspond to masses of potential partial peptides. Two forbidden pairs of vertices (x_1, y_1) and (x_2, y_2) are non-interleaving if the intervals (x_1, y_1) and (x_2, y_2) do not interleave, i.e. one of them is contained inside another. A graph G with a set of forbidden pairs is called proper if every two forbidden pairs of vertices are non-interleaving.

Tandem mass-spectrometry peptide sequencing problem corresponds to anti-symmetric longest path problem in a proper graph. We submit that there exists an efficient algorithm for anti-symmetric longest path problem in a proper graph.

We assume that there are no two vertices u and v in the spectrum graph G such that w(u)+w(v)=w(P); if this happens we shift one of the vertices by a `microscopic' distance ε. We say that edge e={uv} " covers " vertex x when $w(u) < w(P)-w(x) < w(v)$.

We define " combined graph " C(G) as a graph having a path that corresponds to a path in spectrum graph that is folded in the middle. The vertices of the combined graph are pairs (e,x) such that edge e covers vertex x. There are two distinguished vertices in the combined graph. An initial vertex corresponds to pair $(v_{\{initial\}},v_{\{final\}})$ and a final vertex $(v_{\{P/2\}},v_{\{P/2\}})$ corresponds to a folding point of the spectrum graph.

Two vertices $(e_1=\{u_1, v_1\},x_1)$ and $(e_2 = \{u_2,v_2\},x_2)$ are connected by a (directed) edge when $x_1=u_2$ and $x_2=v_1$ or when $e_1=e_2$ and there is an edge x_1x_2 in the spectrum graph G. The rules for the initial and final vertices of the combined graph are slightly different. There is an edge from $(v_{\{initial\}},v_{\{final\}})$ to $(\{uv\},x)$ when $u = v_{\{initial\}}$ and there is edge $xv_{\{final\}}$ in G or when $v = v_{\{final\}}$ and $v_{\{initial\}}x \in G$. Vertex $(\{uv\},x)$ is connected with final vertex of combined graph C(G) whenever x=u or x=v.

The major property of the combined graph we use in our algorithm is that forbidden pairs will get close to each other.

The following establishes the locality of forbidden pairs. Let the maximal distance m between offsets from Δ be smaller then the weight of the smallest amino acid. If x_1, x_2 be a forbidden pair and if p is a path in G(S) from (e_1,x_1) to (e_2,x_2) then p consists of one edge. A proof follows. Every path p with length more than 1 contains an

edge of spectrum graph, therefore the distance between x_1 and x_2 is more than the weight of an amino acid. Therefore x_1 and x_2 cannot be generated from the same peak of the spectrum and pair (x_1,x_2) is not a forbidden pair.

The algorithm for creating a graph without forbidden pairs follows:

- generate spectrum graph G
- generate combined graph C(G)
- for every forbidden pair x_1, x_2 remove edges
- (e_1,x_1) \to (e_2,x_2) from C(G)
- find the shortest path p from initial to final vertex in C(G)
- recover the shortest path without forbidden pairs in G from p.

Although the proof of this theorem is complicated the resulting algorithm will be rather fast and practical. Also sometime we can gain a reasonable solution by searching for paths in opposite direction, starting from the vertex $v_{\{final\}}$ and ending in vertex $v_{\{initial\}}$.

To make the spectrum graph approach work, all vertices that correspond to ion-types of a partial peptides P_i have to be merged into a single vertex corresponding to P_i. Since $\varepsilon = 0.45$ the distance between $s_i + \delta_i$ and $s_j + \delta_j$ is bounded by $0.45 + 0.45 = 0.9$. This, rather large error range, presents a serious problem for merging vertices in the spectrum graph.

We use a greedy algorithm to merge vertices. At every step we find the closest vertices, u (generated from peak s) and v (generated from t) and merge them. The weight of new vertex will be the weighted average (i(s) u+i(t) v)/(i(s)+i(t)) of weights of u and v. We repeat merging until all vertices are at least ε apart for a given precision ε. Note that in the later stages of this merging algorithm we might merge vertices that are already created by merging, in such case the new weight of the vertex is the weighted average of three (or even more) weights of original vertices. The greedy algorithm for merging provides satisfying results for most spectra. However there are cases when the algorithm does not merge vertices related to the same partial peptide or merges vertices that are not associated with the same partial peptide. The doubly charged ions frequently cause problems since their error range is actually twice larger comparing to

error ranges of singly charged ion types. Unfortunately, the greedy merging algorithm described above allows only the uniform error range.

When different error ranges are needed we can proceed in the hierarchical manner. Instead of generating all vertices at once and merging them afterwards we generate only vertices corresponding only to the most significant ion types and merge those vertices using greedy merging algorithm. In the next step we generate the vertices for third most significant ion type and then merge new vertices with the old one. We continue until all vertices are generated and merged. Analysis of histograms in Fig. show frequences of offsets between most frequent ion types leads to a conclusion from that error range in vertex merging can be chosen 0.5 rather than 0.9 as one would expect (data are not shown).

Whenever the distance between two vertices u and v in the spectrum graph is equal to the mass of an amino acid a we connected u and v with an edge and labeled it a. In the last sections we redefined vertices and allowed their weights to be non-integer. In a more realistic approach we join vertices u and v we require that the mass of an amino acid a is approximately equal to the distance between the two vertices, i.e.- $\varepsilon < |v-u| - m(a) < \varepsilon$ for error range ε. To determine the appropriate value for ε we check the peaks (say s and t) corresponding to the same type ions of partial peptides P_i, P{i+1 (say a is the last amino acid of P{i+1 not present in P_i). Analysis of the histograms of offsets $|m(t)-s|-m(a)$ for all such pairs of peaks s and t. The analysis of implies that $\varepsilon=0.5$ is an appropriate choice for error range in defining edges of spectrum graph (data are not shown).

We have observed, that when creating spectrum graph it sometime happens that due to the merging procedure the weights of appropriate vertices are off more $\varepsilon = 0.5$ even when there are corresponding peaks with difference within 0.5 of the amino acid mass. Since such vertices are not connected by an edge, we are at risk of loosing important edges in the spectrum graph. To avoid it we introduce bridge edges in the spectrum graph. We connect two vertices u and v either by a (regular) edge with label a if $-\varepsilon < |v-u|-m(a) <\varepsilon$ or by a bridge edge if there are peaks $s,t \in S$ and ion type $\delta \in \Delta$

such that $-\varepsilon<|s-t|-m(a)<\varepsilon$ and vertex $s+\delta$ was merged into u and vertex $t+\delta$ was merged into v.

A peak of a spectrum is actually a mass/charge (m/z) ratio of the corresponding ion. Up to this point we worked as if z = 1 and assumed m/z of the peak is the same as the mass of the corresponding ion. However, some Mass-spectrometers are capable of producing ions with charge 2 or even more, in this case observed mass is half (third,...) of the ion's actual mass.

We analyze doubly charged ions in the same manner as we did ordinary ions by treating them as a `new' ion type. We investigate offset frequency function $H^{+2}(x,S)$ where offsets are given by $m(P_i)-2s_j$. The analysis of the corresponding offset frequency function demonstrates that the only two significant multiple-charged ion types are y^{+2} and $y^{+2} - H_2O$ (1).

We use simple alignment of spectra to compute parent masses. If $S=\{s_1,...,s_m\}$ is the spectrum of a peptide P $\overline{S} = \{\overline{s}_1,...,\overline{s}_m\}$ then the reflection of S is a spectrum $\overline{S} = \{\overline{s}_1,...,\overline{s}_m\}$ such that $\overline{S}_i = m(P)-s_i-d$, where d = m(y-ion)-m(b-ion) is the difference of offsets of y-ions and b-ions. Note that if a spectrum S contains a peak s that corresponds to a b-ion of a partial peptide P_i and peak t that corresponds to a y-ion of P_i^- then $\overline{S} = t$ and therefore spectra S and $\overline{S}$ have a common element. For correct m(P) we should see good alignment between peaks corresponding to b-ions in S and peaks corresponding to y-ions in $\overline{S}$ (and vice versa because of symmetry).

We use this observation to devise an algorithm for computing the parent mass. For a spectrum $S = \{s_1,...,s_m\}$ and a number x we define $\overline{S}(x) = \{\overline{s}_1,...,\overline{s}_m\}$ where $si = x - \delta_i - d$. Spectra S and $\overline{S}$ may have some peaks in common just by chance, for a `random' mass x the number of peaks in common is approximately $\frac{m(P)}{d^2(S)} \approx 0.5.\ \frac{m(p)}{2601}$ (for thresholded spectra with d(S)=51). It implies that two random spectra have approximately 0.5 peaks in common. However for x = m(P) spectra S and $\overline{S}$ tend to have more peaks in common due to the alignment between b-ions and y-ions. Since the condition that both P_i and P_{i-} ions are present in the spectra is satisfied in 45% of cases

(average number of aligned peaks is 6.4) we are able to devise the following combinatorial approach to estimate m(P).

Let $c(S, \overline{S}(x))$ be the number of peaks $s_i \in S$ and $\overline{s}_i \in \overline{S}(x)$ such that $|s_i-\overline{s}_j|<\varepsilon$, where ε is given precision. The value of x that maximizes $c(S, \overline{S}(x))$ then would be an appropriate choice for parent mass. Should there be many choices for x, we can select one that minimizes the sum of distances $|s_i-\overline{s}_j|$ of the aligned peaks $s_i \in S$ and $\overline{s}_i \in \overline{S}$.

This approach significantly improves the accuracy of the parent mass determination. This approach can similarly be used to correct a mis-assignment of the parent mass/charge value resulting from an incorrect charge assignment.

What is claimed:

1. A method for generating a partial amino acid sequence for a fragmented peptide using mass spectroscopy, the method comprising the steps of:

producing a mass spectrum for said fragmented peptide and transforming said mass spectrum into a spectrum graph whereby each peak in the mass spectrum is represented in said spectrum graph as a plurality of peaks which are offset by values related to a family of possible peptide ion types, and

generating said partial amino acid sequence by deriving the longest path in said spectrum graph that does not include vertices for both a N-terminus and C-terminus fragment ion type representing a single peak in said mass spectrum.

2. A method for determining the precursor mass/charge of a fragmented peptide from a mass spectrum of said fragmented peptide, comprising:

a) reflecting the mass spectrum about the axis of a proposed precursor mass/charge, taking into account the mass/charge offset between a pair of symmetric N-terminal and C-terminal fragment ion types; and

b) aligning the original mass spectrum and the reflected mass spectrum while varying the mass/charge offset necessary to optimize alignment; and

c) adjusting a proposed precursor mass/charge by the mass/charge offset to provide optimal alignment of the original and reflected mass spectra.

3. A method for generating a partial amino acid sequence for a fragmented peptide using mass spectroscopy, the method comprising the steps of:

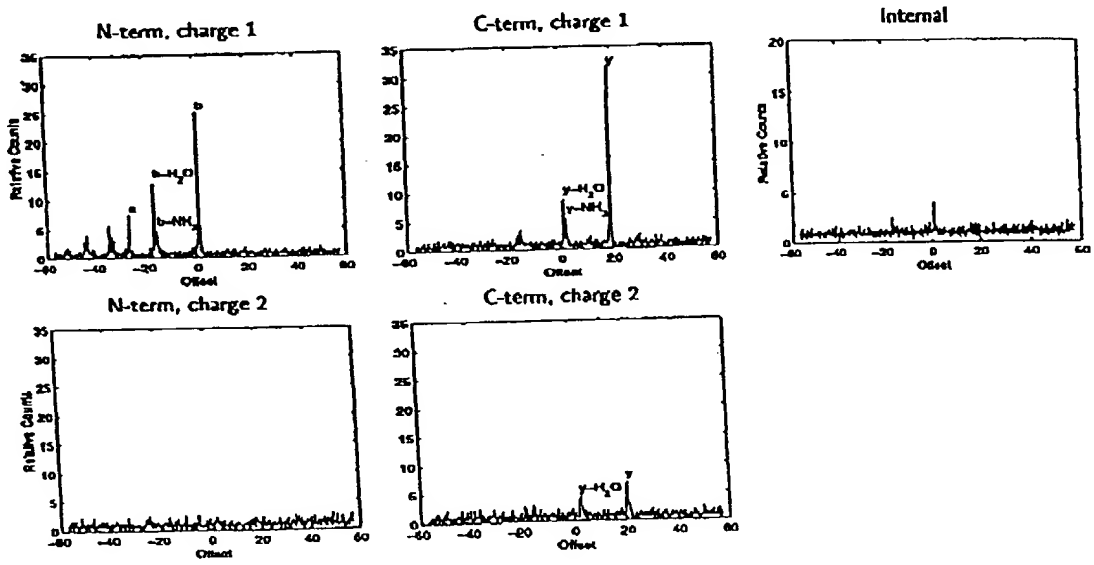
producing a mass spectrum for said fragmented peptide and transforming said mass spectrum into a spectrum graph whereby each peak in the mass spectrum is represented in said spectrum graph as a plurality of peaks which are offset by values related to a family of possible peptide ion types,

generate a combined graph from said spectrum graph,

removing each edge from said combined graph representing an edge between a forbidden pair,

finding the longest path in said combined graph without forbidden pairs,

generating said partial amino acid sequence from said longest path in said combined graph without forbidden pairs.



Plots of the offset frequency functions. Horizontal axes represent offsets between peaks in spectra and masses of partial peptide molecules. Vertical axes represent normalized offset counts with 1 being the average count. Offset increment=0.2. The only significant offsets for internal ions correspond to b and b–H_2O ions. The only significant offsets for doubly charged ions correspond to y and y–H_2O ions

FIG. 1

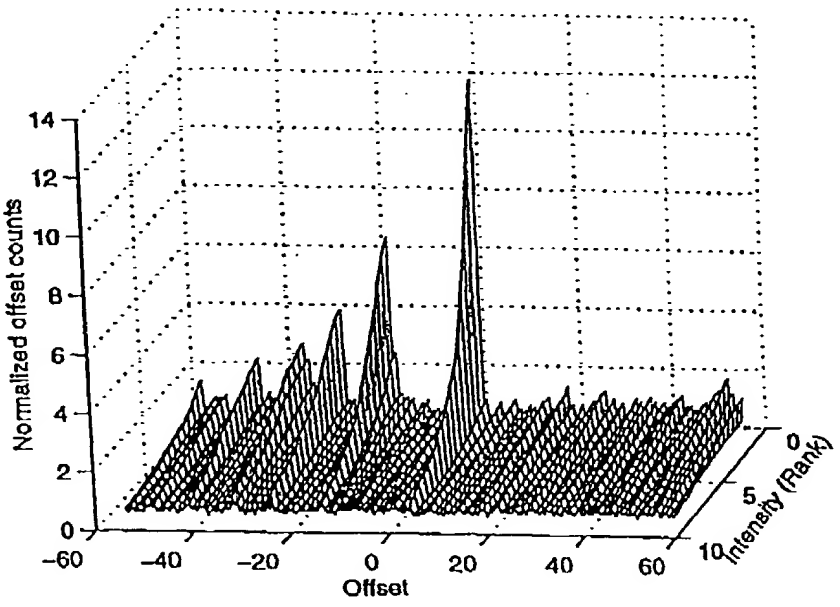


Figure 2: Offset frequency function $H(x, r)$ for N-terminal ion types with intensities ranked r and below.

FIG. 2

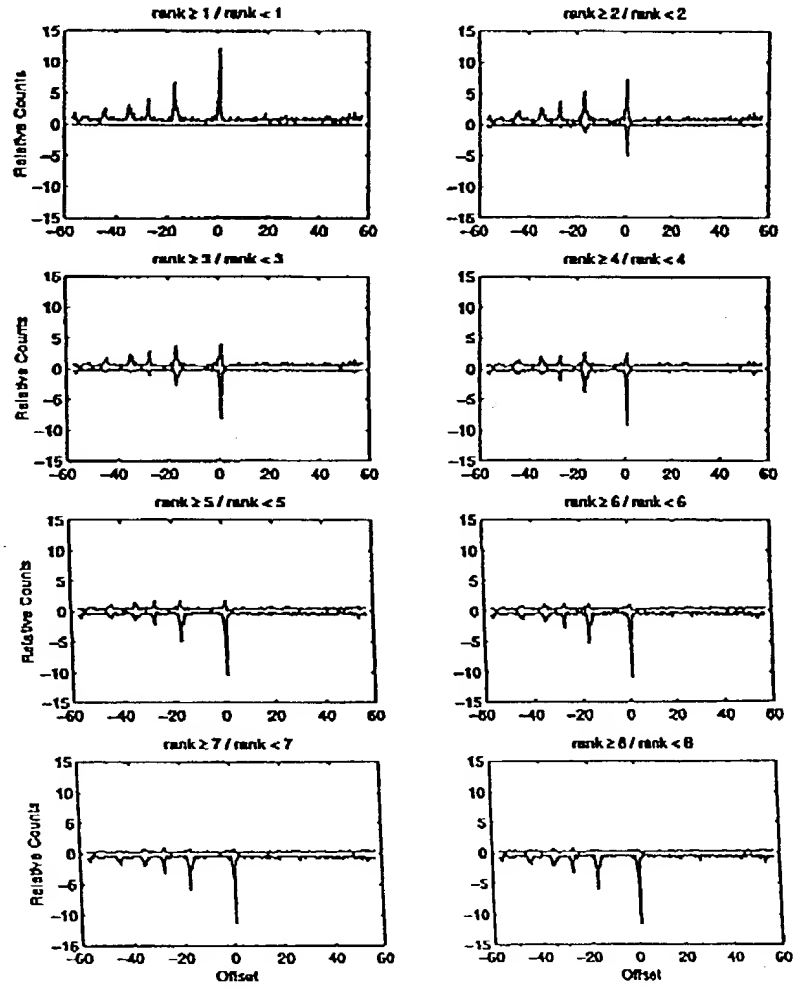
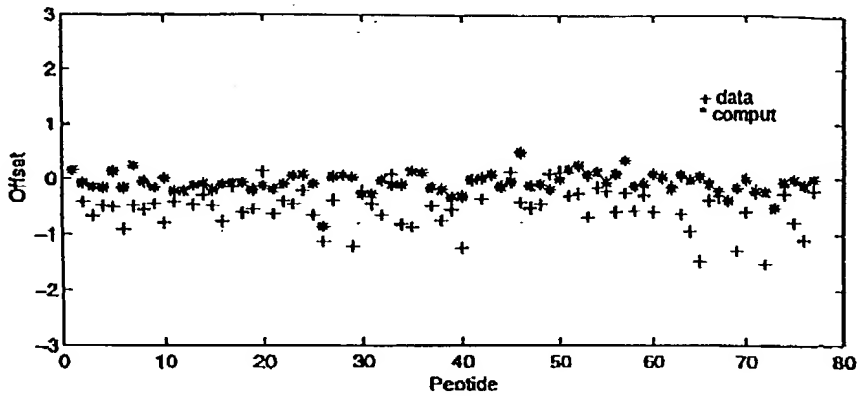


Figure 3: Offset frequency function $H(\pm)$ for N-terminal ion types depending on rank (the size of the bins for computing ranks is (parent mass)/100). Plots shows offset frequency function for ions with rank at least i (upper parts) and with rank less then i (lower parts).

FIG. 3



The offsets between experimentally observed parent masses and $m(P)$ are marked by '+'. The offsets between combinatorially computed parent masses and $m(P)$ are marked by '*'. The average error for parent mass computed by our algorithm is 0.0766 (standard deviation 0.1844), while for observed parent mass it is 0.4743 (standard deviation 0.3732).

FIG. 4